Internship Report

Corona Virus Forecast
And its impact on air pollution in India .

BY –Rajarshi Chowdhuri

TABLE OF CONTENTS

1. THE CORONAVIRUS [Covid-19]

1.1. Introduction to Coronavirus

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus.. Most people infected with the COVID-19 virus will experience mild to moderate respiratory illness and recover without requiring special treatment. Older people, and those with underlying medical problems like cardiovascular disease, diabetes, chronic respiratory disease, and cancer are more likely to develop serious illness.

The best way to prevent and slow down transmission is be well informed about the COVID-19 virus, the disease it causes and how it spreads. Protect yourself and others from infection by washing your hands or using an alcohol based rub frequently and not touching your face

1.2. Symptoms

The main symptoms include:

- Fever
- Coughing
- Shortness of breath
- Trouble breathing
- Fatigue
- Chills, sometimes with shaking
- Body aches
- Headache
- Sore throat
- Loss of smell or taste
- Nausea
- Diarrhoea

The virus can lead to pneumonia, respiratory failure, septic shock, and death. Many COVID-19 complications may be caused by a condition known as cytokine release syndrome or a cytokine storm. This is when an infection triggers your immune system to flood your bloodstream with inflammatory proteins called cytokines. They can kill tissue and damage your organs.

How does the coronavirus spread?

SARS-CoV-2, the virus, mainly spreads from person to person.. Most of the time, it spreads when a sick person coughs or sneezes. They can spray droplets as far as 6 feet away. If you breathe them in or swallow them, the virus can get into your body. Some people who have the virus don't have symptoms, but they can still spread the virus.

You can also get the virus from touching a surface or object the virus is on, then touching your mouth, nose, or possibly your eyes.

1.3. Objective:

To analyze data from various sources on the global pandemic of coronavirus and see its effect/impact on air pollution.

1.4. Approach

I have done this entire project keeping India in mind. The results may vary from country to country and should not be taken as concrete. Also there were lot of data insufficiencies so had to replace and drop certain attributes of the different datasets used here.

1.5. Datasets Used

1. COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University

2. COVID19 Global Forecasting (Week 4) dataset

3. Air Quality Data in India (2015 - 2020) Central Pollution Control Board.

The enitire python notebook can be seen from this link. Click here

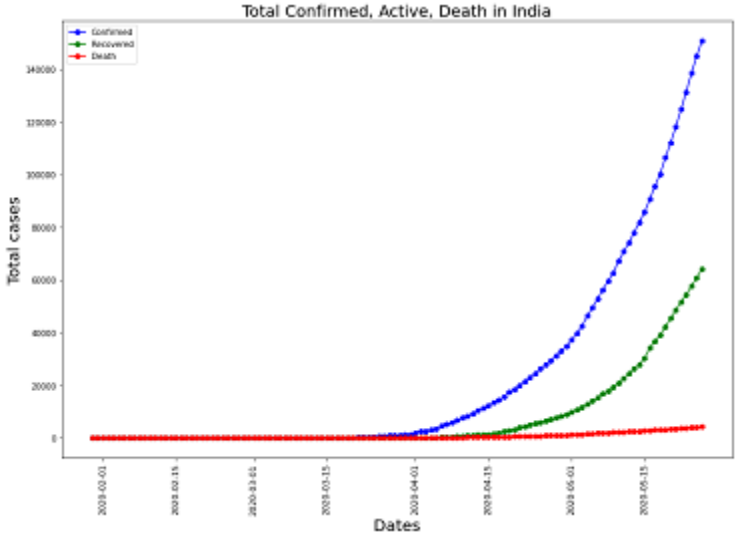
The Story of COVID-19 in India

The COVID-19 pandemic is the defining global health crisis of our time and the greatest global humanitarian challenge the world has faced since World War II. The virus has spread widely, and the number of cases is rising daily as governments work to slow its spread. India has moved quickly, implementing a proactive, nationwide, lockdown, with the goal of flattening the curve and using the time to plan and resource responses adequately. India with a huge population count and density in no exception facing any less wrath of this pandemic. Let's dive deep into the data to find out more about it.

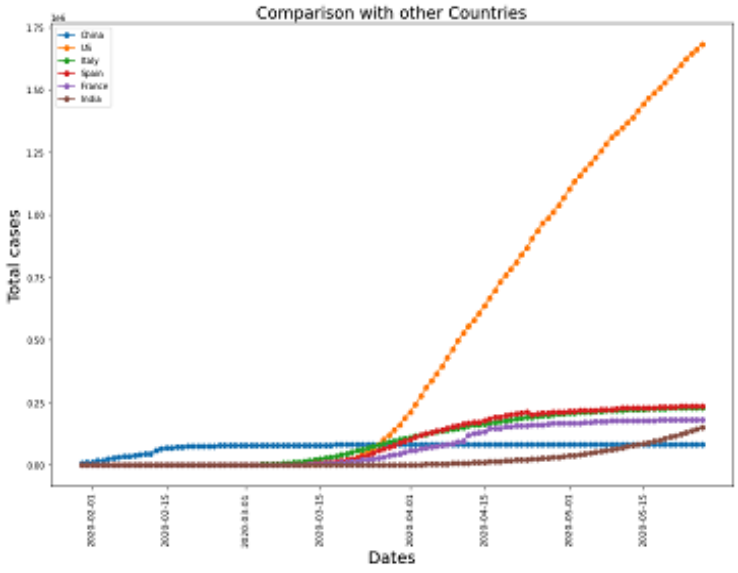
2. ANALYSIS OF DATA

2.1. Current scenario of India

If we check the graph below we can clearly see that the number of active cases has increased from the end week of march and is increasing more or less in an exponential fashion. The recovery line is also increasing steeply which is a good sign but we need to maintain the gap between these two lines as minimum as possible. Deaths due to covid 19 has followed almost a linear path except for this month of May where the line is seen slowly tapering up.



The graph below shows where does India stand among the world. We have taken US,China,France,Italy and Spain.



INFERENCE

.

Though being highly populated the relative confirmed cases of India is low compared to other countries. This could be because of two reasons

- 21 day lockdown imposed by prime minister Narendra Modi (Source : Health Ministry)
- Low testing rate (Source: news18)

According to Health ministry Without lockdown, India would have 2 lakh Covid-19 cases by now. So an early on lockdown may have been beneficial for India ,but situation is degrading and fast. We need to flatten the curve as soon as possible.

The testing rate also need to pick up soon to test and isolate the infected person as well the area or the contacts the person has come in touch with.

On a global scale US is the worst affected country among the lot. Its graph is nowhere close to other countries and has shot up drastically . Comapred to that India is a far better position.

.

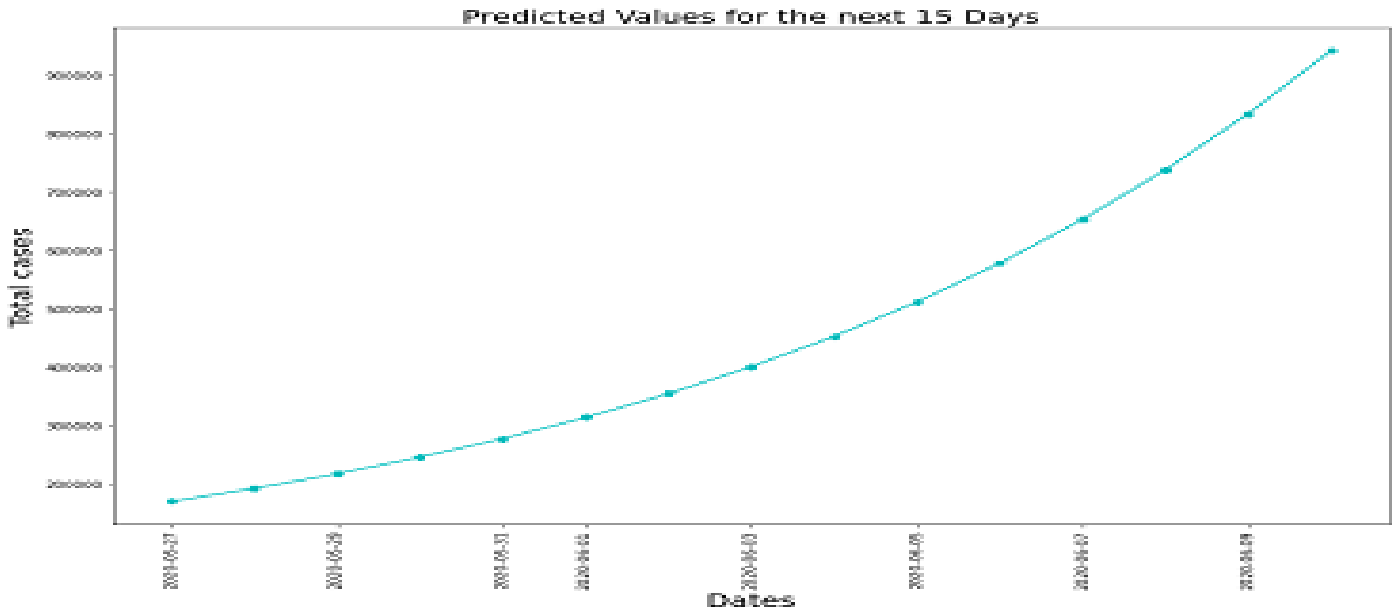
3. PREDICTION

3.1. Prediction using growth factor

We have calculated the growth factor as in the rate at which the numbers are increasing in India. This growth rate can give us a fairly average idea of the coming days for India. The rate though is subjected to the data till May as per the dataset. The rate will be increasing in the coming days for sure keeping in mind the increasing number of people getting infected by this pandemic.
So as per our data the growth factor has come to 1.129 approx

```
Average growth factor 1.129914482388571
```

Using this growth factor let's see how the coming 15 days would be like .Below is the plot of the coming 15 days.

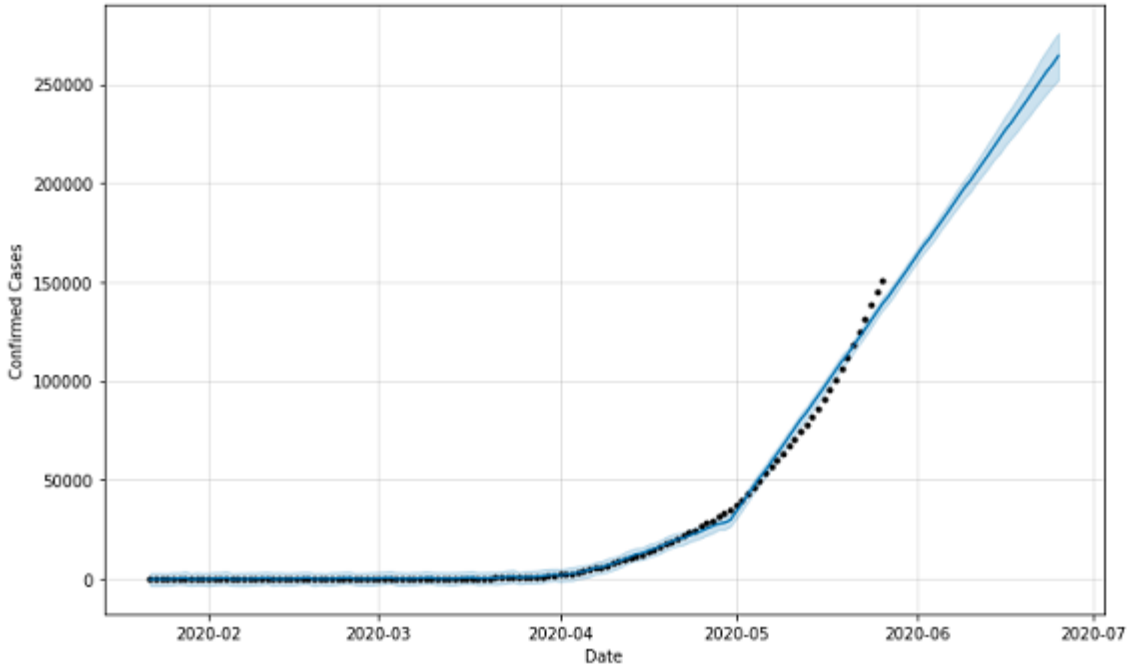


We could see that the graph is increasing exponentially if the average growth factor doesn't decrease. It is important that the growth factor is reduced to flatten the curve.

3.2. Predicting using Facebook Prophet

Facebook Prophet:

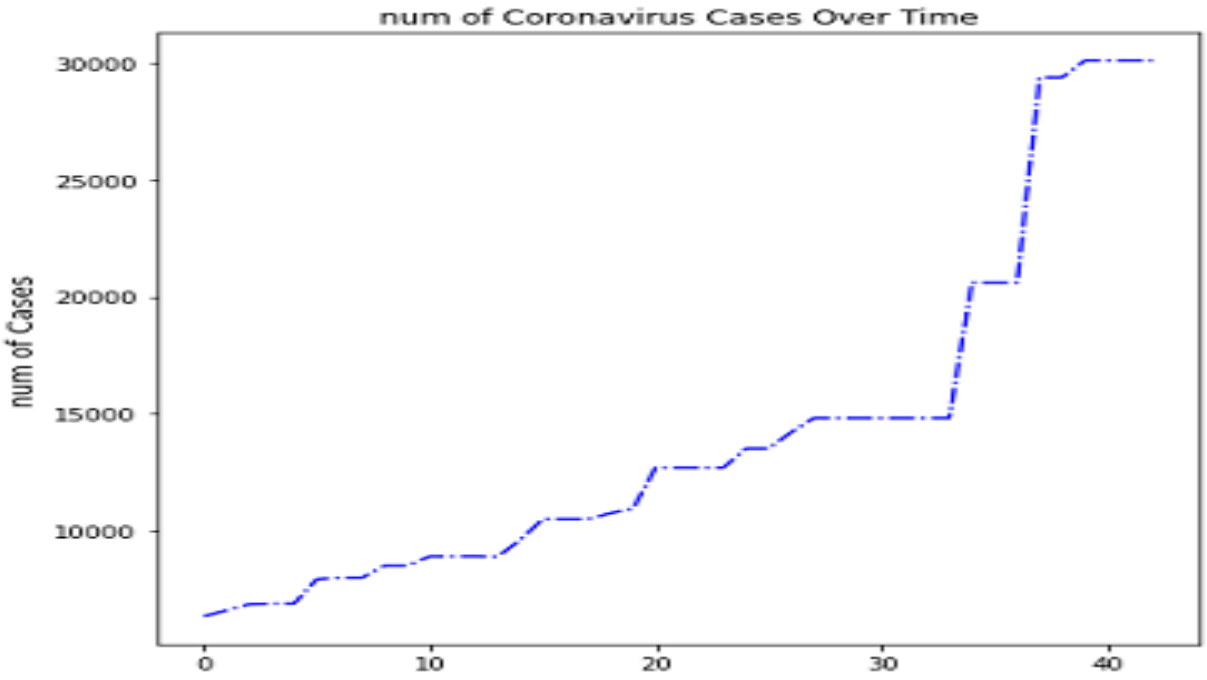
This is a library shaped by Facebook which is very simple to execute with stylishly satisfying visuals. It is additionally a very well made model. Prophet is a method for estimating time series information dependent on an additive model where non-straight patterns are fit with yearly, week after week, and day by day irregularity, in addition to occasion impacts. It works best with time series that have solid occasional impacts and a few periods of verifiable information. Prophet is strong to missing information and moves in the pattern, and ordinarily handles anomalies well.



As projected the graph seems to keep increasing exponentially for the days to come. It has correctly envisioned the growth from the previous records.

3..3. Predicting using XG boost regressor

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now. XGBoost stands for "Extreme Gradient Boosting" and it is an implementation of gradient boosting machines. The XGBoost is a popular supervised machine learning model with characteristics like fast in computation, parallelization, and better performance.



XG boost unlike prophet gives some hope as we can clearly see the graph flattening over time but at the same time the graph shows it has uneven stretches as well .The MSE ,RMSE, MAE are given below.

```
[ ]  mse
     [157416419.57]
[ ]  rmse
     [12546.57]
[ ]  mae
     [4134.78]
```

Between the two Prophet seems to be more logical and practical with the data we have till now that is till May.

4. Effect on Air Pollution

4.1. Overall Objective

We took a look at 24 Indian cities air pollution levels over the years as well as forecast the air pollution levels.The data has been made publicly available by the Central Pollution Control Board portal of Government of India.

There will be two main parts to the project:

1. To compare the various states on the level of pollution for the year 2019.
2. To find trends, seasonality etc for the pollution levels of India as a whole as well as Delhi and forecast it to the future.
3. Checking if the current corona pandemic has affected the pollution levels.

A brief introduction to the calculation of AQI

- The AQI calculation uses 7 measures: PM2.5(Particulate Matter 2.5-micrometer), PM10, SO2, NOx, NH3, CO and O3(ozone).
- For PM2.5, PM10, SO2, NOx and NH3 the average value in last 24-hrs is used with the condition of having at least 16 values.
- For CO and O3 the maximum value in last 8-hrs is used.
- Each measure is converted into a Sub-Index based on pre-defined groups.
- Sometimes measures are not available due to lack of measuring or lack of required data points.
- Final AQI is the maximum Sub-Index with the condition that at least one of PM2 and PM10 should be available and at least three out of the seven should be available.

How is AQI calculated?

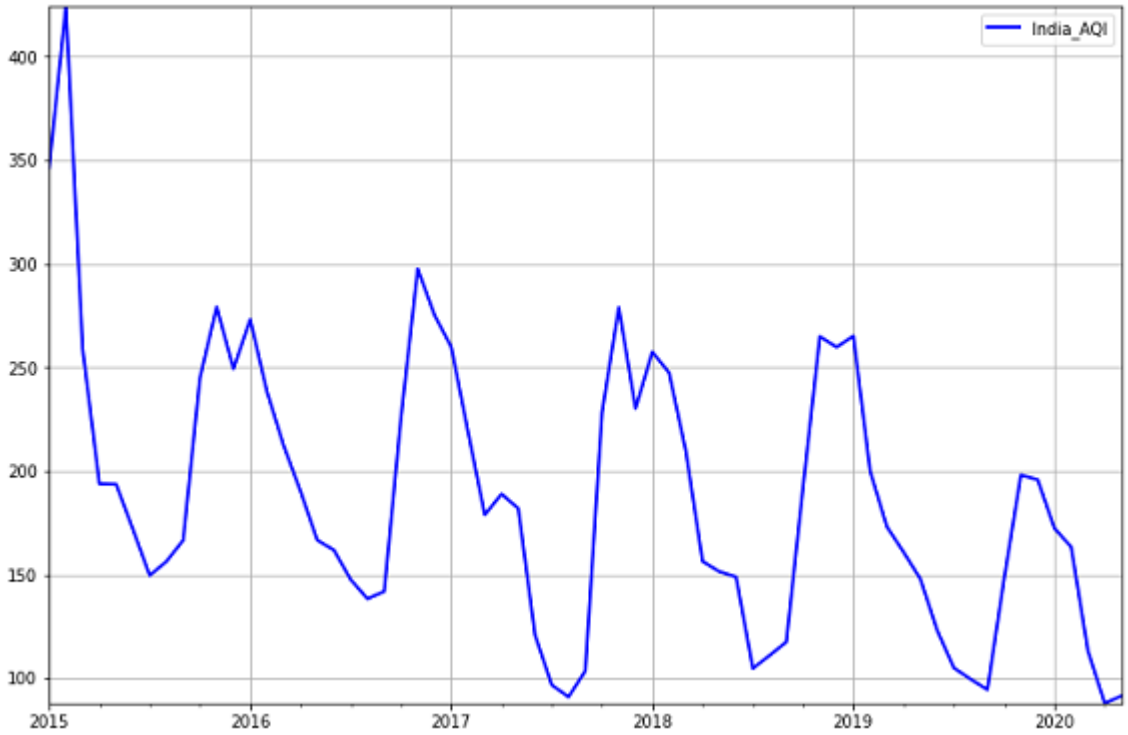
The Sub-indices for individual pollutants at a monitoring location are calculated using its 24-hourly average concentration value (8-hourly in case of CO and O3) and health breakpoint concentration range. The worst sub-index is the AQI for that location.

All the eight toxins may not be observed at all the areas. By and large AQI is determined just if information are accessible for least three contaminations out of which one ought to essentially be either PM2.5 or PM10. Else, information are viewed as inadequate for computing AQI. Essentially, at least 16 hours' information is viewed as important for ascertaining subindex.

The sub-indices for monitored pollutants are calculated and disseminated, even if data are inadequate for determining AQI. The Individual pollutant-wise sub-index will provide air quality status for that pollutant.

The web-based system is designed to provide AQI on real time basis. It is an automated system that captures data from continuous monitoring stations without human intervention, and displays AQI based on running average values (e.g. AQI at 6am on a day will incorporate data from 6am on previous day to the current day).

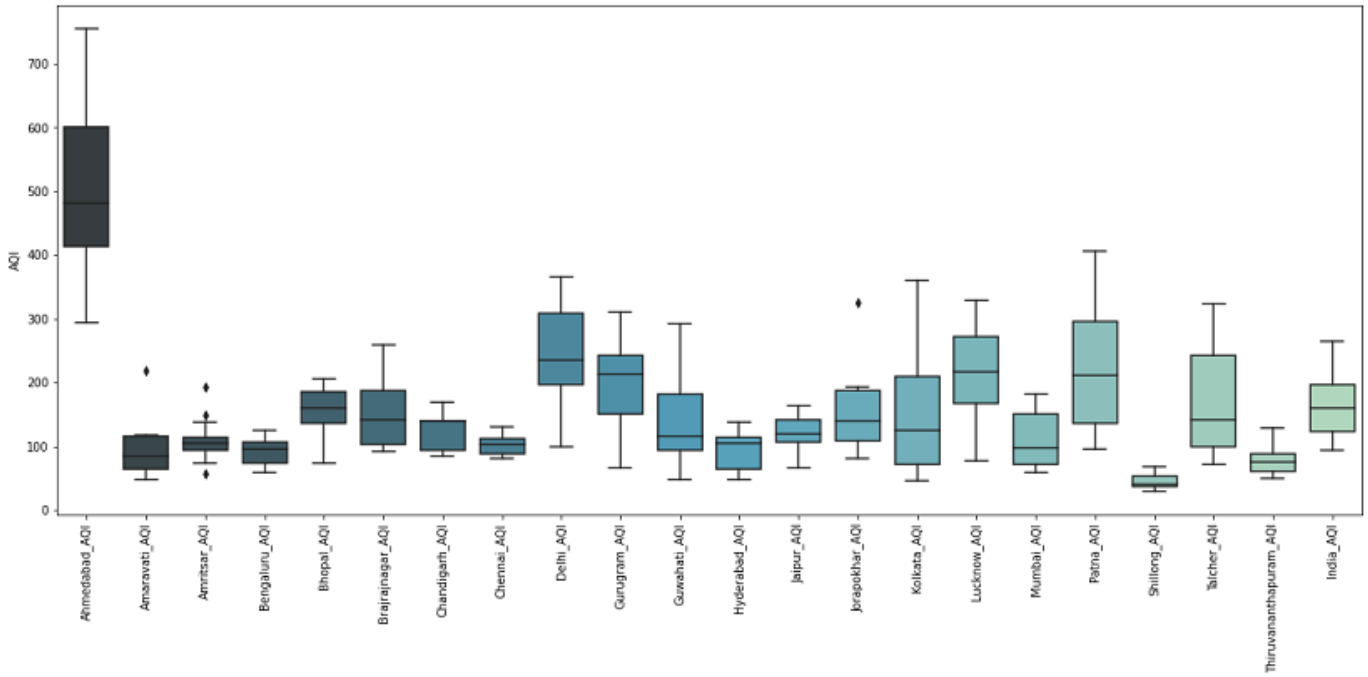
4.2 Checking AQI levels in India over the years



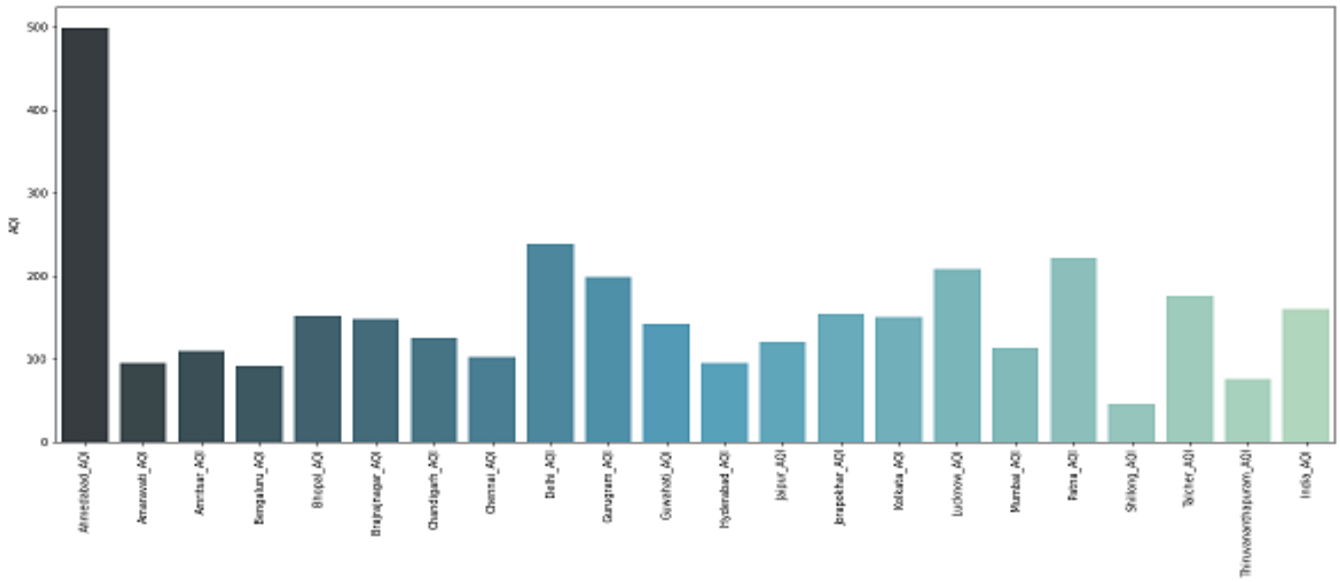
Straight away we can see patterns and trends over the years. There are two highly noticeable patterns. One is the general trend downwards. Over the past 5 years we can see the AQI reducing marginally. Note that this can be a litte misleading, especially due to the 2015 data, as the dataset in the first few observations only comprises of Delhi and Ahmedabad during which have relatively highly pollution compared to the rest of the cities which makes the initial portion of the graph highly exaggerated. Nevertheless we can see a general decline in pollution over the years.

The next pattern thats easily observable is the seasonal component which plays a big role in the pollution of the country. We will discuss further in the 2nd part of our project. One other important point to note is the affect of COVID-19 on India's pollution level. The pollution levels are drastically lower during the year 2020 for the same reason.

Before looking at the means of the AQI values of the cities, we will take a look at the boxplots of the AQI values of the various cities.

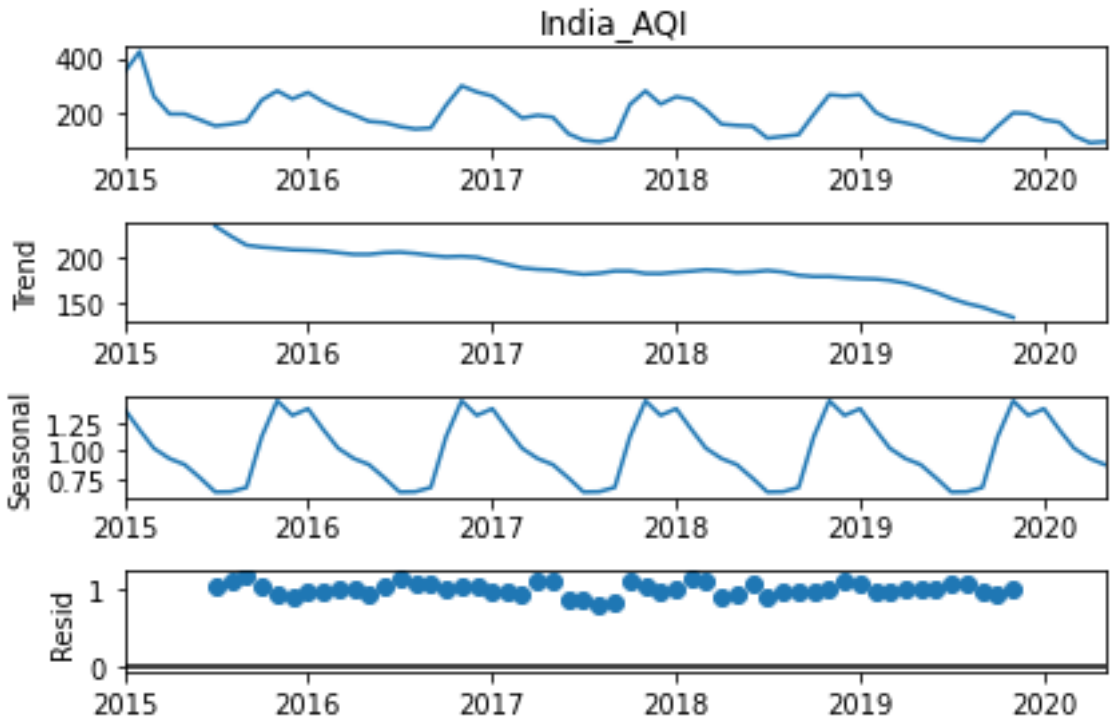


We can see that Ahmedebad has easily the highest values of AQI in the country, followed by Delhi lagging far behind. Let us take a look at the means of the values of AQI for further comparison.



We can see that Ahmedabad and Delhi are the most polluted whereas Shillong is the least followed by Trivandrum. With this we end the comparison and move to the next section of forecasting the values of future AQI for the whole of India.

We will first take a look at the seasonal decompose of the AQI values of india.



As we have discussed earlier ,there is a very clear seasonality, and a less clearer trend. The trend is possibly due to increasing restrictions on pollution by the govt and the last surge downward is clearly due to the recent Covid-19.

We can see that there are two peaks largely, one during October and the during January. And the lowest amount of pollution is around july-September after which there is a sharp increase. Similarly, there is a decrease from January to JulyThis spike in the winters is due to a combination of factorsThe spike is due to factors including Winter aversionseasonal factors such as dust storms, crop fires, burning of solid fuels for heating, and firecracker-related pollution during Diwali, stubble burning etc.During winters the planetary boundary layer is thinner as the cooler air near the earth's surface is dense. The cooler air is trapped under the warm air above that forms a kind of atmospheric 'lid'. This phenomenon is called winter inversion. Since the vertical mixing of air happens only within this layer, the pollutants released lack enough space to disperse in the atmosphere. During summers, pollution levels decrease as the warmer air rises up freely, making the boundary layer thicker, and providing enough space for pollutants to disperse. The same thing happens during winter afternoons, when increased heat brings down pollution slightly.

4.3 Forecasting

We will be using three methods for forecasting values of AQI for India, namely, SARIMA, RNN using LSTM and facebook prophet. It is obviously overkill to be using these three methods however being new to time series I would personally like to explore all three options. Normally for such a small dataset RNN would not be recommended.

Using SARIMA

Autoregressive Integrated Moving Average, or ARIMA, is one of the most widely used forecasting methods for univariate time series data forecasting.Although the method can handle data with a trend, it does not support time series with a seasonal component.An extension to ARIMA that supports the direct modeling of the seasonal component of the series is called SARIMA.

First, we run auto arima to find out the parameters of the model for us. We can manually do it,however, it is much easier for us let the notebook do the work for us.

SARIMAX Results

Dep. Variable:	India_AQI	**No. Observations:**	41
Model:	SARIMAX(1, 1, 1)x(1, 0, 1, 12)	**Log Likelihood**	-204.793
Date:	Wed, 27 May 2020	**AIC**	419.586
Time:	18:45:40	**BIC**	428.030
Sample:	01-01-2015 - 05-01-2018	**HQIC**	422.639

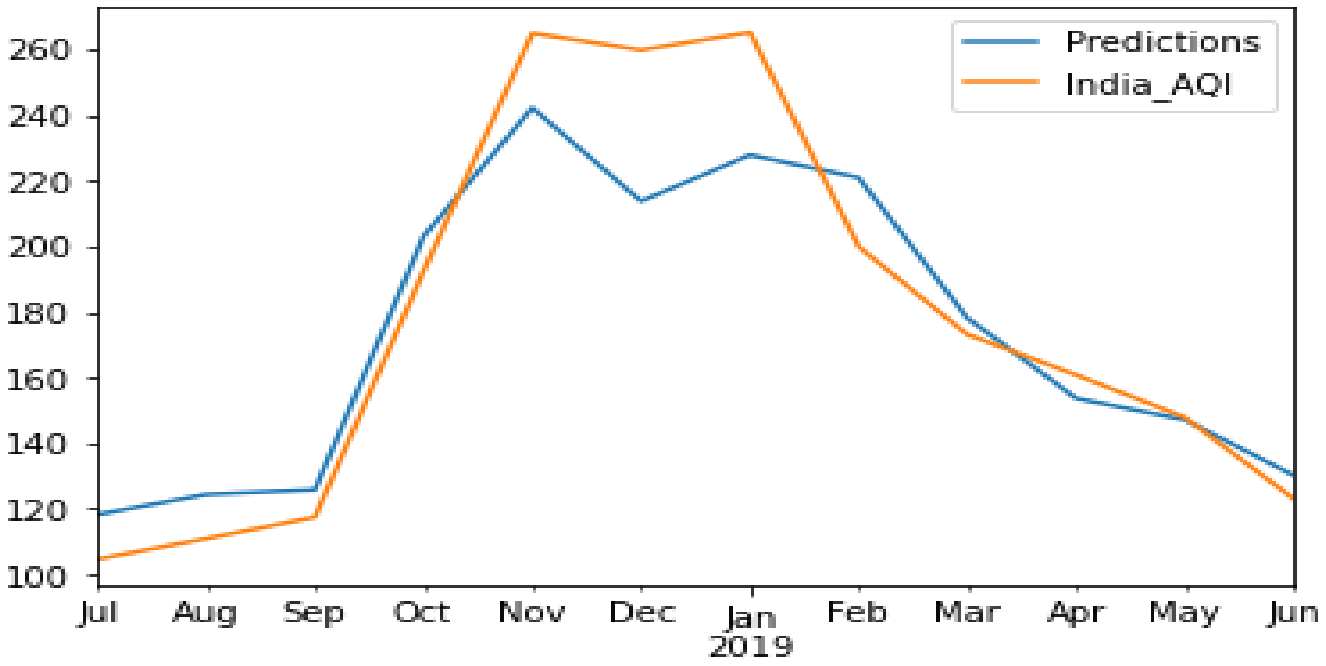
Covariance Type: opg

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6708	0.203	-3.312	0.001	-1.068	-0.274
ma.L1	0.9996	5.867	0.170	0.865	-10.500	12.499
ar.S.L12	0.9310	0.266	3.505	0.000	0.410	1.452
ma.S.L12	-0.6350	0.693	-0.916	0.360	-1.994	0.724
sigma2	1297.9758	7426.082	0.175	0.861	-1.33e+04	1.59e+04

Ljung-Box (Q):	48.19	**Jarque-Bera (JB):**	12.74
Prob(Q):	0.15	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.29	**Skew:**	-0.66
Prob(H) (two-sided):	0.04	**Kurtosis:**	5.43

We have found the optimal parameters for the SARIMAX model is (1,1,1)x(1,0,1,12)

Our next step is to forecast using this model into the future. However, since we do not have information regarding future values, we will split the data into a training data and testing data and try to predict 1 year into the future. We will use the years 2015-2018(till june) as our train dataset and July-June the next year as our test dataset. The reason we exclude 2020 is due to the fact that 2020 is an outlier due to covid an we will not get an accurate figure for the prediction. We will also take a look at the predicted values of 2020 for reference. Further, we will predict into the year 2021.

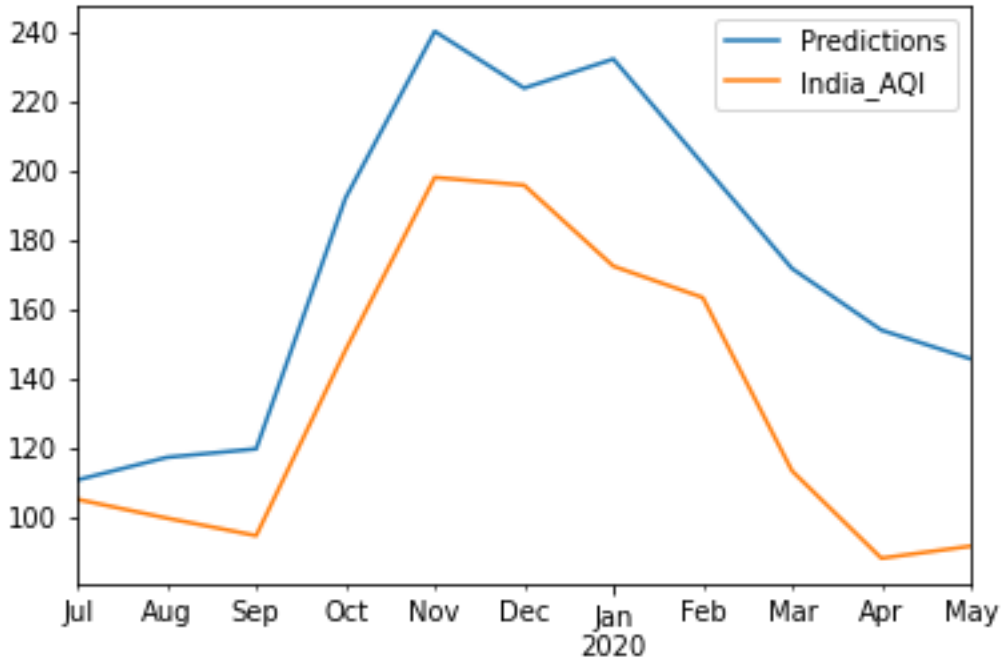


We can see that the predicted values are fairly close to our actual values using SARIMA and hence is quite fascinating how looking at previous values gives us so much insight into future air pollution.However, there is a discrepency at the peak of the graph where our model has not been able to predict with a high accuracy. To obtain the value of error we will be using root mean square error(RMSE) for comparison between the models.

```
RMSE =  20.69607069599271
Mean AQI 176.68763588530942
```

We have got an RMSE value of approximately 21, which is quite alright, we can approximately judge the scale of error by comparing with the mean values of AQI which is 177, so the error is approximately 1/9 of the actual values.

Next we will try predicting the AQI values for the year 2019-2020(July-May)



As expected, the predicted values are much higher than the actual value as we can see from the graphs. Let us take a look at the error value.

```
RMSE =  43.95748314220997
Mean AQI 133.55768465497644
```

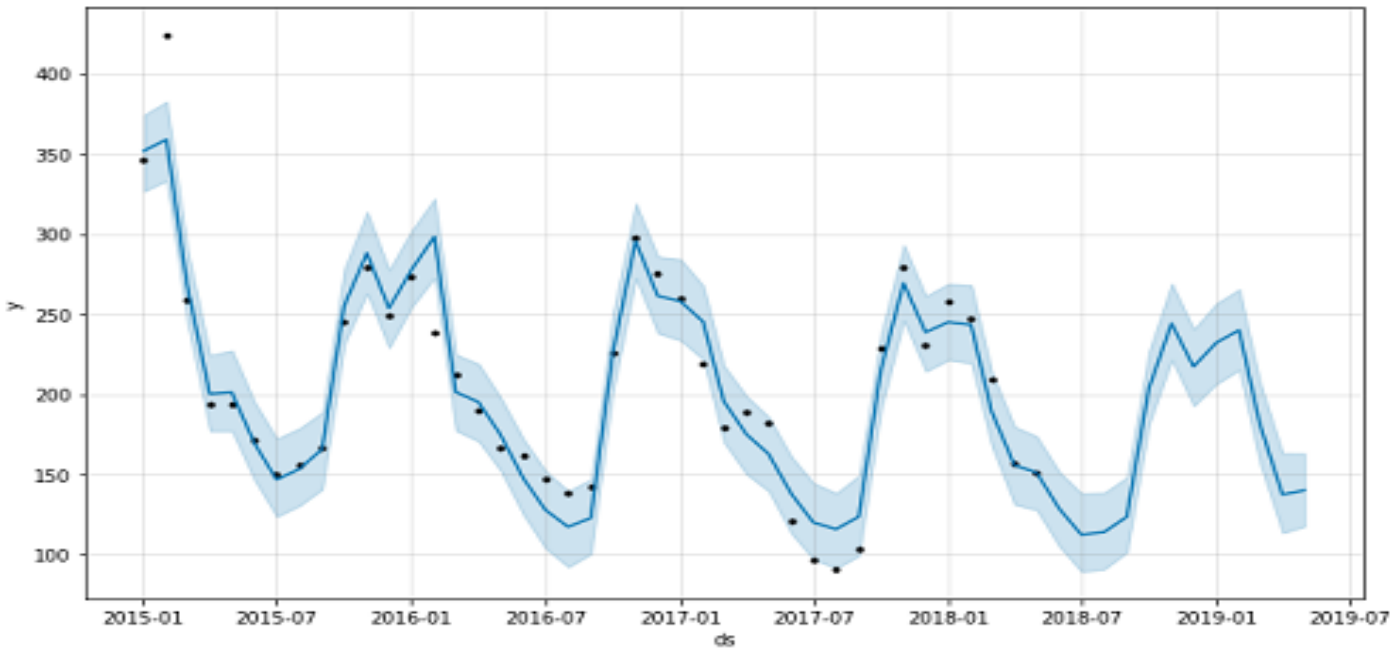
The error value is much higher than earlier for obvious reason and hence we can see that predicting for the year 2020 is not going to yeild accurate results due to the Covid-19.

Next we will take a look into forecasting into the unknown, i.e. 2020-2021. This poses a problem, as if we predict including 2020 data, we are bound to get an innacurate prediction for next year simply due to the fact that 2020 is an outlier.However, if we remove 2020 from our dataset and predict from 2019 till 2021 we are left with wrong predictions for sure and considering that covid-19 could have further lasting effects we will predict poorly. We will choose to include 2020 as well for this predicition. We could compare the values next year.

We can see the predictions plotted in continuation with 2020 and one thing we note is the highly optimistic prediction. That is purely due to the fact that 2020 is such an outlier, chances are, the pollution levels will follow the trend pre 2020 which would mean a bump in the AQI levels unless the country decides to keep the restrictions etc as is which is highly unlikely. We can always get a more accurate prediction skipping 2020.

Using Facebook Prophet:

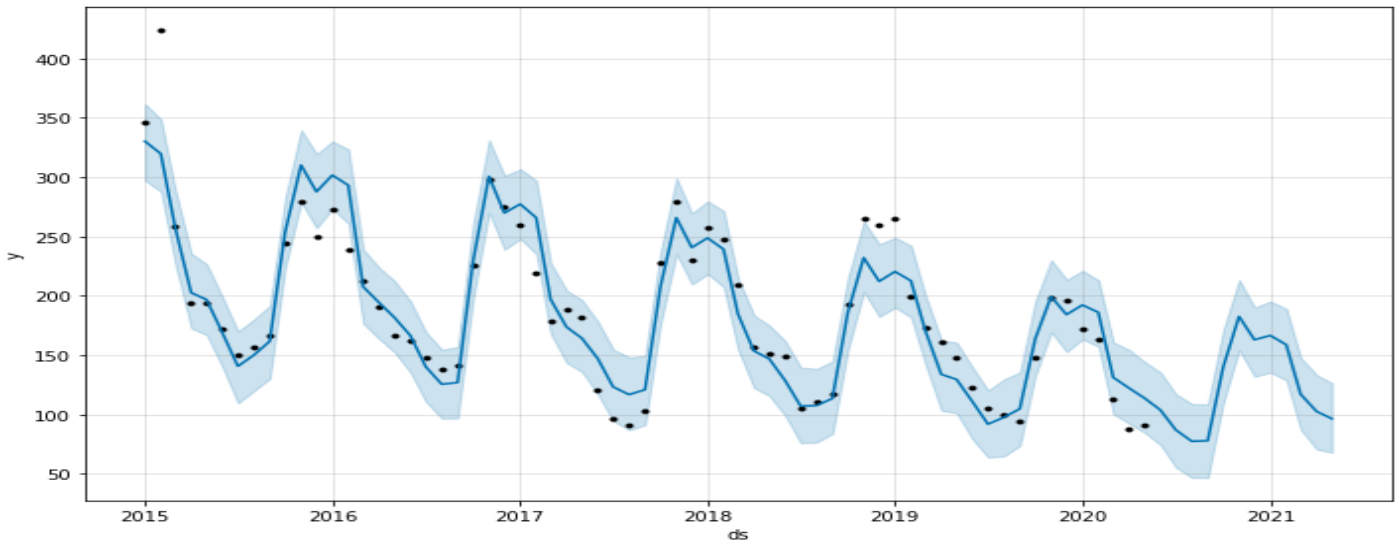
As seen before we use the prophet algorithm again to predict the pollution levels.




```
RMSE =  22.814309479104242
Mean AQI 178.83105077960582
```

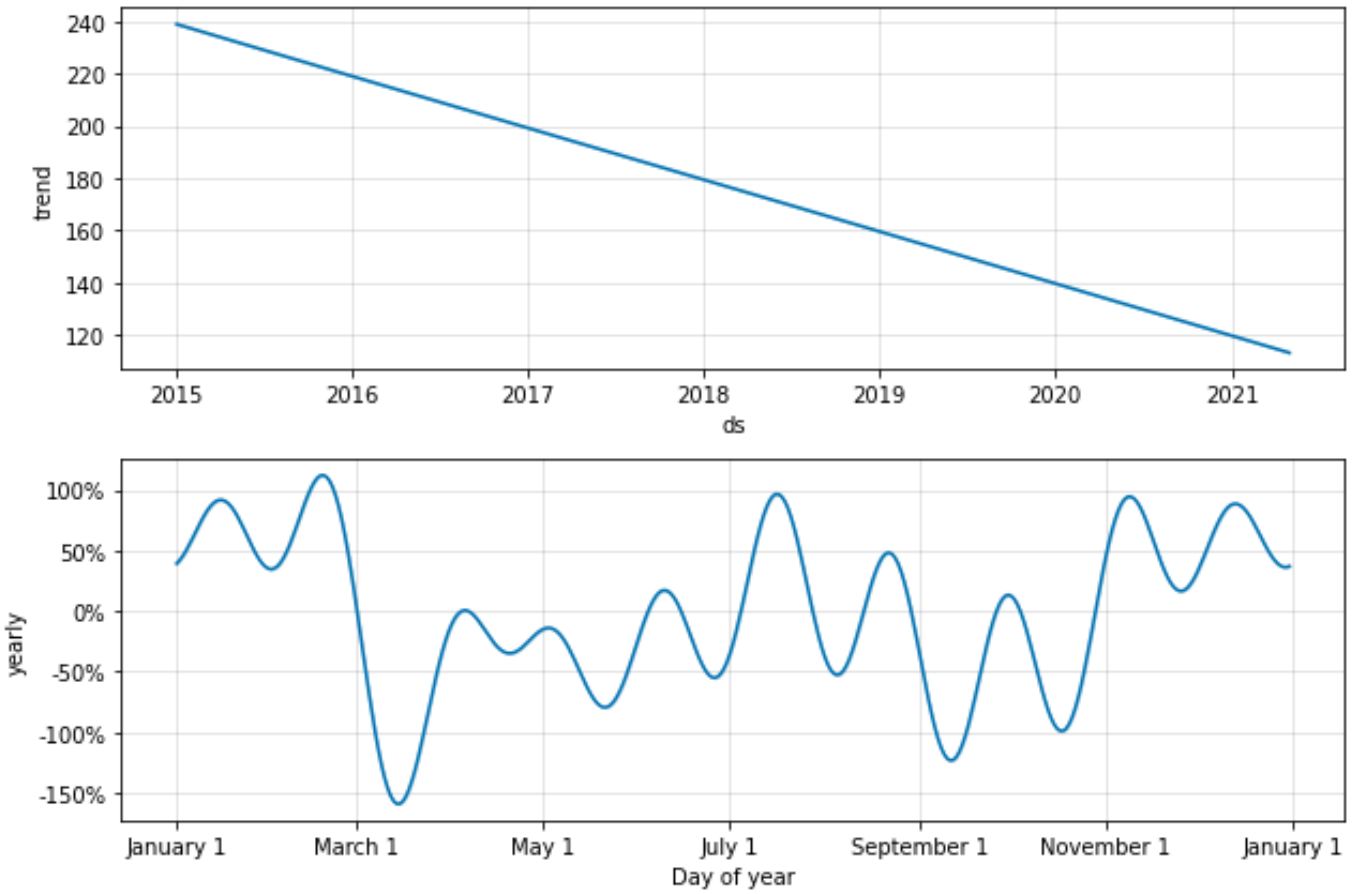
The RMSE value is a lot higher than that of SARIMA, this could be indicative that Prophet weighs long past data with more weight. Perhaps the model could be further tweaked to get better results.

Lastly, let us predict into the future using the Prophet Library.



This is the final plot into the future. Surprisingly, this model looks like it could be a predictor better than SARIMA for this case considering the trend has not been too greatly altered for due to covid. This further seems to indicate that prophet seems to be placing more emphasis on past values as compared to SARIMA. Note that we can retrieve the predicted values from the forecast object.

Finally let us take a look at the components of our data found by prophet:

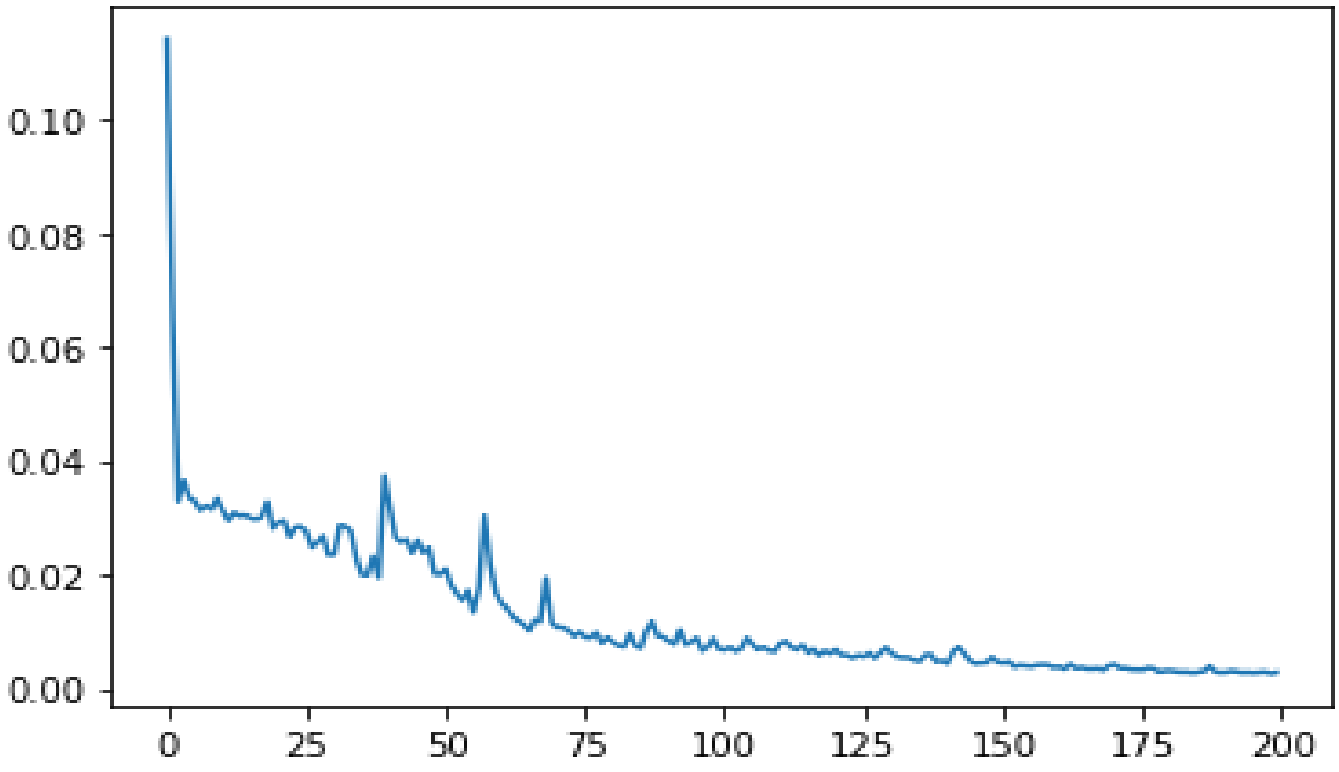


The above diagram does give us a much clearer indepth idea of the trend and seasonal component of our data. With this we have come to an end to the prediction of AQI values using prophet

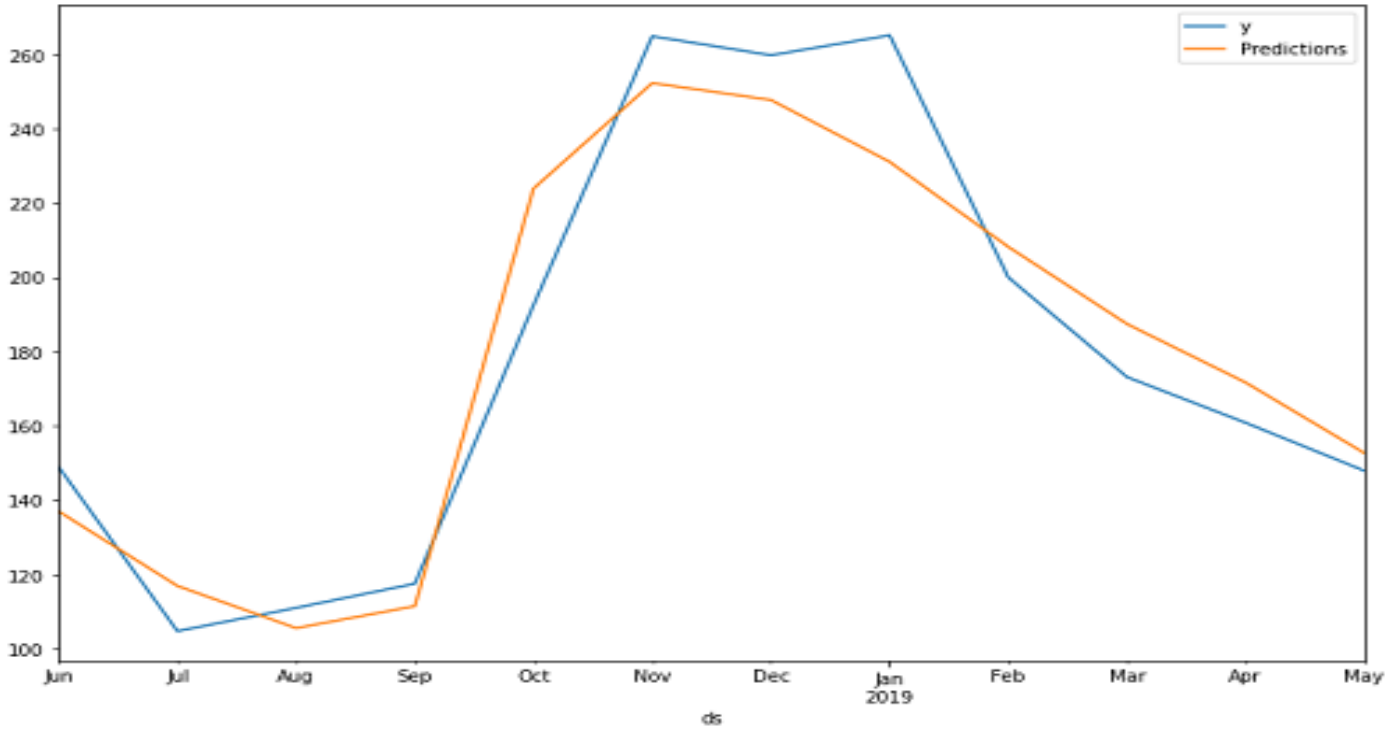
Using Recurring Neural Networks(RNN):

For the last forecast we will use RNN which is a type of Neural Network that is used for time based/frequency based/ memory based data like text data, speech, time series etc. We will be using a particular cell type LSTM(Long short term memory).LSTM networks are particularly meant to keep particular information for a longer term as compared to regular RNN's. As all Neural Networks, RNN's works best with a huge amount of data. RNN is a black box method, which means there is little transparency in the model and how it trains. Another major disadvantage is the high complexity of hyperparameters. Hence RNN's should preferably used as last resort.

The plot below shows how the values of the loss reduces as each epoch gets over is shown below.



Using Lstm to predict the values gives more or less a satisfactory result.




```
RMSE =  16.370820649533652
India_AQI= 188.32307331319683
```

The RMSE value is lower than what we had predicted with the above two models even with our limited dataset. We know that the lower the RMSE more better is the model . So based on this we can say this model worked quite well with respect to others.

Forecasting into the future with RNN:

We will use the same model but with the entire dataset now and predict one year into the future

Again, like that with SARIMA, we can see that the prediction is highly optimistic due to COVID-19 which can possibly be better by removing the year 2020 and predicting two years in using data pre 2020.

5. CONCLUSION

The novel coronavirus disease (COVID-19) pandemic brings upon a cruel choice to the world: the society as well as the economy. It has shown the weaknesses and strengths of every country and has taught us a series of lessons. The first question is, have we accepted the problem? Every year, millions of people die of air, water or land pollution.

Close to 1.2 million people die due to of air pollution alone. In Delhi and the National Capital Region (NCR) itself, millions are affected. We can imagine the size and horror of the situation given the fact that 15 of the world's 20 top-most polluted cities are located in India and there are large number of Indian cities that do not go by the World Health Organization air quality standards.

It is a well-established that air pollution is a silent killer that affects our lungs and causes serious health issues, especially, in the elderly and children. Still, we haven't accepted the problem. Our Union and state governments are not ready to accept the problem unless the National Green Tribunal (NGT) or the Supreme Court issues some order to them or against them. Indeed, it is the judiciary that is driving the environmental debate in India.For example, the Environment Pollution (Prevention and Control) Authority, a Supreme Court-mandated body, is tasked with taking various measures to tackle air pollution in the NCR.

The application of environmental norms in the most polluting coal-based sector seems impossible without the intervention of the Supreme Court. Second, have we communicated the seriousness of the problem? No. How will we communicate when our government itself says 'No Indian study has shown any correlation between pollution and the shortening of lifespan'.

This brief Exploratory Data Analysis and forecasting here shows just the tip of the iceberg of the grave situation we are in. More work needs to be done and newer insights might come up with more data coming in in the coming days of this pandemic. The only silver lining here is that the forecast shows this pandemic has decreased the pollution levels to a great extent such that the world has started to heal and can be seen practically by various news reports and activities around the world.

But the question remains to what extent? Will this pave way to newer policies and changes that will help to lower the pollution levels or it is going to stay the same as lockdown ends …only time can tell.

6. BIBLIOGRAPHY

Websites :

1. https://edition.cnn.com/2020/03/31/asia/coronavirus-lockdown-impact-pollution-india-intl-hnk/index.html
2. https://www.downtoearth.org.in/blog/pollution/lessons-from-covid-19-on-reducing-india-s-environmental-pollution-70891
3. https://en.wikipedia.org/wiki/Coronavirus
4. Others..

Papers:

1. https://facebook.github.io/prophet/
2. https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/43905.pdf
3. https://iopscience.iop.org/article/10.1088/1757-899X/394/5/052024/pdf
4. Others..

7. END NOTES….

Till the day the vaccine or any medicine for the coronavirus comes into place there is no shadow of doubt that this pandemic will continue. All we can do is practice social distancing and minimize going out to reduce interaction which eventually will lead to flattening the curve. Wash your hands, stay at home, save yourself, save your family. Stay safe.